DIGITAL SIGNAL PROCESSING 1, 215–230 (1991)

Methods for Waveform Interpolation in Speech Coding

W. Bastiaan Kleijn and Wolfgang Granzow

AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974-2070

1. INTRODUCTION

Most speech coding algorithms operating at rates of around 8 kb/s attempt to reproduce the original speech waveform. Their efficiency of reproducing the waveform is obtained by using models which exploit knowledge of the generation of the speech signal. In contrast, most coders operating at rates of around 2.4 kb/s are completely parametric, usually transmitting parameters describing the pitch and the spectral envelope at regular intervals. However, because of model inadequacies, the quality of reconstruction of current parametric methods never reaches that of the original signal, even at high bit rates.

In this paper, a new method which is positioned between the waveform coders and the parametric coders is presented. It is based on the assumption that, for voiced speech, a perceptually accurate speech signal can be reconstructed from a description of the waveform of a single, representative pitch cycle per interval of 20–30 ms. Figure 1 shows the smooth evolution of the shape of the pitch cycle, which is typical for voiced speech signals. We will show how such a signal can be reconstructed by interpolating *prototype* pitch cycles between the updates. The prototype-waveform interpolation (PWI) method retains the natural quality typical of coders which encode the entire waveform, but requires a bit rate close to that of the parametric coders.

We discuss PWI methods based on linear prediction (LP). In LP-based speech coders, the signal is reconstructed from knowledge of the predictor coefficients and a description of the excitation signal. Of the existing LP-based algorithms, the code-excited linear-prediction (CELP) algorithm [1] and the LP vocoder [2] are examples of waveform and parametric coders, respectively.

In the simplest form of CELP the speech waveform is described by time-varying LP filter coefficients and a filter excitation consisting of the concatenation of scaled fixed-length vectors from a codebook. To achieve high efficiency during voiced speech, most implementations include a long-term predictor [3], or adaptive codebook [4], to facilitate periodicity of the reconstructed signal. Despite recent improvements [5, 6], inaccurate reproduction of the periodicity remains the main source of perceptual distortion in the current CELP algorithms at rates below 6 kb/s.

In the LP-based vocoders the voiced speech signal is modeled by a single pulse per pitch cycle. Because of excessive periodicity, this often leads to a buzzy character of the reconstructed speech. Recent work has shown that the speech quality can be improved significantly by adding more information about the evolving waveform shape. Using a cluster of pulses for each pitch cycle, with blockwise shape adaptation, in combination with a smoothly varying overall gain produced good results [7]. Alternatively, good-quality voiced speech can be obtained at rates of around 3 kb/s by careful placement of the single-pulse locations [8, 9]. Although significantly improved over the LP-based vocoders, and similar in quality to 4.8 kb/s CELP, such single-pulse excited (SPE) speech coders still suffer from some buzziness.

Both the CELP and the SPE methods attempt to reproduce the original waveform by using a (spectrally weighted) signal-to-noise ratio (SNR) of the reconstructed speech signal as a criterion to determine the excitation sequence. However, maintaining the periodicity of the original speech signal is important for its perceptual quality, and maximization of the SNR often leads to a nonoptimal degree of periodicity. Thus, it was found in both the CELP [6] and the SPE coders [9] that improved speech quality can be obtained by increasing the periodicity, despite an associated reduction in SNR.

1051-2004/91 $1.50

FIG. 1. A 14-ms segment of a voiced speech signal, uttered by a female speaker and band-limited to 4 kHz.

Maintaining a smooth pitch track and the correct degree of periodicity is fundamental to time-scale modification of speech, making it useful to look at these methods from a speech coding viewpoint. Recently, excellent results in time-scale modification were obtained with methods which are pitch synchronous [10, 11]. The basic idea behind these procedures is to add pitch cycles for slowing down the speech rate, and to eliminate pitch cycles for increasing the speech rate. The success of these methods implies that, in order to maintain a good speech quality, transmission of information during each pitch cycle is not essential, but that it is important to maintain a high correlation between neighboring pitch-cycle waveforms, and to provide a good description of these waveforms. In other words, these time-scaling techniques suggest that less frequent updates of the waveform, combined with interpolation, may result in an efficient encoding.

Recent developments in speech coding by sinusoidal reconstruction also point toward the importance of maintaining the correlation between the waveform of successive pitch cycles. Traditionally, coders using sinusoidal reconstruction have been sensitive to reverberation. Recently, it was shown that reverberation could be eliminated by maintaining phase coherence during successive pitch cycles [12]. In informal experiments, we have found that the reconstructed speech signal has a reverberant quality when it is modified by an added component which changes rapidly over time, but which itself has a harmonic structure similar to that of the original speech signal. Thus, a reverberant character is caused by an added component which is correlated from one pitch-cycle to the next, while a noisy character (such as that of basic CELP) is caused by an added component that is not correlated between successive pitch cycles. In both noisy and reverberant speech signals, it is the dynamics of the pitch-cycle waveform that is disturbed.

These results suggest that, in voiced speech, it is important to maintain the original dynamics of the pitch-cycle waveform. In PWI this is accomplished by interpolation in combination with other features. The PWI procedure can be seen as a generalization of the original vocoder concept, which transmits not only the pitch period and the LP filter specification but also a prototype excitation waveform at each update. Similar to CELP, an analysis-by-synthesis method is used to quantize the excitation waveform. However, in the PWI method a single pitch cycle is quantized every 20–30 ms, whereas in CELP the speech waveform is quantized on a frame-by-frame basis. The excitation waveform and the filter parameters are interpolated independently between updates. Alternatively, the method can be interpreted as a simple vocoder with single pulses exciting a time-dependent pole-zero filter, where the excitation waveform provides the coefficients of the all-zero filter.

The PWI method avoids the common problems which are caused by incorrect dynamics of the pitch-cycle waveform. Noise, buzziness, and reverberation can be controlled because the speech signal is reconstructed on a pitch-cycle by pitch-cycle basis. In cases where noise is present in the original signal, such as in speech with a breathy character, this can be modeled by adding white noise to the excitation signal. Reverberation and noise are controlled by maintaining the correlations between sequential pitch-cycle waveforms similar to that of the original signal. By maintaining the level of periodicity of the original speech signal, excellent sounding reconstructed speech can be obtained at bit rates of 2.5 to 4 kb/s.

We proceed as follows. In the next section, we describe the principles of the PWI in more detail. In Section 3, we provide the practical details for implementing the method. In Section 4 we show how the method performs. We conclude in Section 5 with a summary of the main distinguishing features of the PWI method.

2. PRINCIPLES OF THE PWI METHOD

2.1. *Separation of Spectral Envelope and Spectral Fine Structure*

The two major features of a speech spectrum are the formant structure (the spectral envelope) and the fine structure. As a rough model, the formant structure is determined by the vocal tract shape, while the fine structure is determined by the vocal cords. We consider the formant structure as being independent of the fine structure and assume that these two features evolve separately over time. For good results, these features must be interpolated separately in the PWI method. This can be done by deconvolving the speech signal into an excitation signal which is white, but maintains the spectral fine structure (pitch), and a description of the formant structure.

It is convenient to model the formants with standard LP techniques. The spectral envelope can be encoded and interpolated using well-known procedures. For this reason, we have chosen to describe the prototype pitch cycle with a set of LP filter coefficients on the one hand, and the pitch and a description of the excitation waveform over one pitch cycle on the other hand. The excitation waveform is constructed by interpolation of the prototype excitation waveforms. The reconstructed speech signal is obtained by filtering this excitation waveform with a time-varying all-pole filter, similarly as in other LP-based techniques. Subsection 2.2 describes the basic construction of the excitation waveform by interpolation, while Subsection 2.3 describes the explicit control of noise and reverberation.

2.2. *Interpolation of the Excitation Waveform*

As a first step in the LP-based PWI reconstruction method, the excitation is computed. It is obtained from interpolation of the prototype excitation waveforms, each describing one pitch cycle. To obtain good quality for the reconstructed speech signal, its pitch contour must be sufficiently smooth, and the correlations between adjacent pitch cycles must be sufficiently high. In this section we will discuss several methods satisfying these conditions. We start out with describing two blockwise interpolation procedures and then we discuss a continuous interpolation method. For convenience of notation, we will describe the excitation as a continuous, rather than a sampled, signal.

2.2.1. Blockwise interpolation with fixed time scale. Let $p(k)$ be the pitch period of an arbitrary pitch cycle k, and let the current interpolation interval start in the center of pitch cycle $k = 0$, with pitch period $p(0)$, and end in the center of pitch cycle $k = K$, with pitch period $p(K)$. We interpolate the pitch period linearly with the pitch-cycle index k:

$$p(k) = \frac{K-k}{K}p(0) + \frac{k}{K}p(K), \qquad k = 0, 1, \ldots, K. \quad (1)$$

The time locations, t_k, of the centers of each of the pitch periods, k, are obtained by simply adding the pitch periods of prior pitch cycles:

$$t_k = t_0 + \frac{1}{2}\sum_{l=1}^{k}(p(l) + p(l-1)) = t_0 + p(0)k + (p(K) - p(0))\frac{k^2}{2K}, \qquad k = 0, 1, \ldots, K. \quad (2)$$

Note that this implies that, for the entire interpolation interval,

$$t_K - t_0 = \frac{K}{2}(p(0) + p(K)). \quad (3)$$

This simple relationship holds if one interpolates linearly in the pitch-cycle index k; linear interpolation in time leads to less elegant results, but performs equally well.

Let us denote the prototype excitation waveforms associated with pitch cycle 0 and K by $v(0, \tau)$ and $v(K, \tau)$, respectively, where $v(0, \tau)$ is defined on the interval $[-\frac{1}{2}p(0), \frac{1}{2}p(0))$ and $v(K, \tau)$ is defined on the interval $[-\frac{1}{2}p(K), \frac{1}{2}p(K))$. In the first interpolation method we define extended, centered prototype excitation waveforms by zero-padding,

$$\tilde{u}(m, \tau) = \begin{cases} v(m, \tau), & -\frac{1}{2}p(m) \leq \tau < \frac{1}{2}p(m), \\ 0, & \text{elsewhere}, \end{cases} \quad (4)$$

where m takes the values 0 and K. The tilde indicates that the extended prototypes $\tilde{u}(m, \tau)$ are centered around the origin. This centering does not mean that the features (e.g., the pitch pulse) of successive prototype excitation waveforms are aligned. The prototype waveform function $u(m, \tau)$ denotes the waveform after proper alignment with the previous prototype excitation waveform. The alignment procedure consists of shifting $\tilde{u}(K, \tau)$ over a distance ξ_K so as to minimize a distortion measure $D(u(0, \tau), \tilde{u}(K, \tau - \xi_K))$:

$$\xi_K = \underset{\xi_K}{\operatorname{argmin}}\, D(u(0, \tau), \tilde{u}(K, \tau - \xi_K)). \quad (5)$$

We then have

$$u(K, \tau) = \tilde{u}(K, \tau - \xi_K). \quad (6)$$

To prevent divergence of the offset ξ_K, it is convenient to recenter the prototype waveform associated with pitch cycle 0 of the interpolation interval, prior to interpolation. That is, the t_0 of the next interpolation interval corresponds to $t_K + \xi_K$ of the present interpolation interval. The definition of the offset for pitch cycle 0,

$$\xi_0 = 0, \quad (7)$$

will be used from here on.

In the present blockwise interpolation method, the intermediate excitation pitch-cycle waveforms $u(1, \tau), \ldots, u(K-1, \tau)$ are obtained from linear interpolation with the pitch-cycle index:

$$u(k,\tau) = \frac{K-k}{K}\,u(0,\tau) + \frac{k}{K}\,u(K,\tau), \qquad k = 0, 1, \ldots, K. \quad (8)$$

To obtain the actual pitch-cycle excitation waveform $v(k,\tau)$, for pitch cycle k, the function $u(k,\tau)$ is truncated to the proper length $p(k)$ of Eq. (1):

$$v(k,\tau) = u\left(k, \tau + \frac{k}{K}\xi_K\right), \qquad -\tfrac{1}{2}p(k) \leq \tau < \tfrac{1}{2}p(k), \quad k = 0, 1, \ldots, K. \quad (9)$$

The excitation waveform $x(t)$ is obtained by concatenation of the truncated waveforms, starting at t_0:

$$x(t) = \sum_{k=0}^{K} u(k, t - t_k)\,\Xi\left(p(k) + \frac{\xi_K}{K}, t - t_k - \frac{k}{K}\xi_K\right), \qquad t_0 \leq t < t_K + \xi_K, \quad (10)$$

where $\Xi(a,t)$ is a window function:

$$\Xi(a,t) = \begin{cases} 1, & -\tfrac{1}{2}a \leq t < \tfrac{1}{2}a, \\ 0, & \text{elsewhere.} \end{cases} \quad (11)$$

The length of the windows for the individual pitch cycles within Eq. (10) does not equal the pitch period, because of the alignment shift ξ_K. The simple dependence of the window function length on ξ_K results again from the fact that the interpolation is linear with the pitch-cycle index. If the truncation operation can be neglected, then the procedure of zero-padding followed by linear interpolation corresponds to linear interpolation of the complex spectrum of the prototype excitations, on a pitch-cycle by pitch-cycle basis.

The interpolation procedure is illustrated in Fig. 2. In this example the left prototype excitation waveform is a centered, band-limited impulse, and the right prototype waveform is an offset band-limited impulse (of lower cut-off frequency). The right prototype has a pitch period 50% larger than that of the left prototype. The left prototype requires a large amount of zero-padding during the interpolation, as is seen in Fig. 2b. However, the right prototype also requires zero-padding, because of the offset of the impulse. In general, the zero-padding can give rise to discontinuities at the block boundaries, and within the blocks at the ending of the prototype waveforms.

Because of the discontinuities, it is essential for the blockwise methods that areas of high energy in the excitation waveform (such as pitch pulses) are known, such that the endpoints of the prototype waveforms can be located where they have the least impact. The discontinuities can be eliminated with a simple modification of the present formalism. The pitch periods $v_0(\tau)$ and $v_K(\tau)$ can be defined over an interval somewhat longer than one pitch period, and the square window $\Xi(a,t)$ can be replaced by an asymmetric tapered window extending to the center of the pitch cycles before and after the present pitch cycle. Then Eq. (10) describes an overlap-add procedure, which results in a smooth transition between adjacent interpolated pitch-cycle excitation waveforms. However, to obtain proper alignment, it is good policy to locate the centers of the prototype waveform near pitch pulses, even when overlap-add procedures are used.

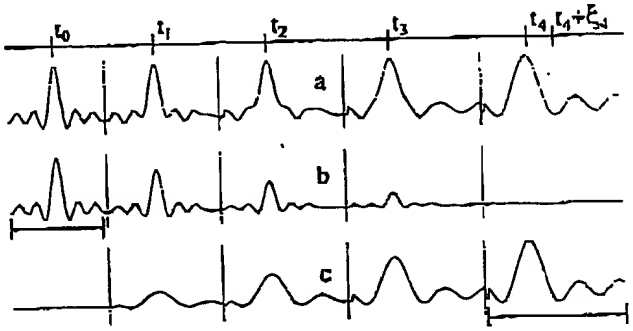


FIG. 2. Blockwise interpolation with fixed time scale: (a) the interpolated excitation waveform, (b) the contribution from the left prototype excitation waveform, and (c) the contribution of the right prototype excitation waveform. The vertical lines indicate the block boundaries and the horizontal bars delineate the prototype excitation waveforms.

The implementation of the blockwise interpolation with fixed time scale proceeds as follows. Initially, the synthesizer is provided with $p(0), p(K), v(0,\tau), v(K,\tau), t_0$, and a "desired" endpoint t_{update}. First the prototypes are aligned according to Eqs. (5) and (6). By substituting t_{update} for t_K in Eq. (3), a noninteger value is obtained for K, which is rounded up to the next integer. Then the actual t_K is computed using Eq. (3). From this the locations of all the intermediate pitch cycles are computed using Eq. (2). Finally, the entire excitation function is computed by using Eqs. (8) and (10). After recentering the last prototype excitation waveform of the present interval, the interpolation over the next interval can proceed.

2.2.2. *Blockwise interpolation with time scaling.* In the previous blockwise interpolation procedure we zero-padded the excitation waveform, aligned the futuremost prototype with the previous prototype, and then interpolated. In the present blockwise interpolation procedure, the prototype waveform, $v(k,\tau)$, is considered to be one cycle of a periodic function $u(k,\tau)$,

$$u(k,\tau) \approx v(k, \mathrm{mod}(\tau + \tfrac{1}{2}p(k), p(k)) - \tfrac{1}{2}p(k)), \quad (12)$$

where the modulus function $\mathrm{mod}(a,b) \equiv a - b\,\mathrm{int}(a/b)$ is used. Before alignment and interpolation of the prototype excitation waveforms, their pitch periods, $p(0)$ and $p(K)$, are normalized to unity. The normalized (dimensionless) time scale is denoted by $\bar{\tau}$. Since the periodic function goes through one cycle for a unity increase in $\bar{\tau}$, we identify $2\pi\bar{\tau}$ as the pitch-cycle phase (the phase of the fundamental harmonic). To obtain the aligned waveform $u(K, p(K)\bar{\tau}) = \bar{u}(K, p(K)(\bar{\tau} - \bar{\xi}_K))$ of the time-normalized excitation waveform we minimize a distortion criterion:

$$\bar{\xi}_K = \underset{\bar{\xi}_K}{\mathrm{argmin}}\, \mathrm{D}(u(0, p(0)\bar{\tau}), \bar{u}(K, p(K)(\bar{\tau} - \bar{\xi}_K))). \quad (13)$$

The interpolation of the time-normalized waveforms, followed by time denormalization by a factor $p(k)$, results in the interpolated excitation waveform for pitch cycle k:

$$u(k,\tau) = u(k, p(k)\bar{\tau}) = \frac{K-k}{K}\,u(0, p(0)\bar{\tau}) + \frac{k}{K}\,u(K, p(K)\bar{\tau}), \quad k = 0, 1, \ldots, K. \quad (14)$$

The excitation waveform is obtained by concatenation of the denormalized waveforms of the individual pitch cycles according to Eq. (10) with $\xi_K = p(K)\bar{\xi}_K$.

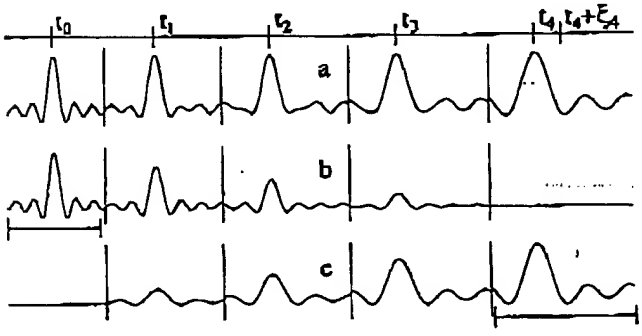


FIG. 3. Blockwise interpolation with time scaling: (a) the interpolated excitation waveform, (b) the contribution from the left prototype excitation waveform, and (c) the contribution of the right prototype excitation waveform. The vertical lines indicate the block boundaries and the horizontal bars delineate the prototype excitation waveforms.

The method is illustrated in Fig. 3. The prototype excitation waveforms are identical to those of Fig. 2. These prototype waveforms make up one pitch cycle of a smooth, periodic function (Section 3 describes methods for extracting prototype excitation waveforms with this property). Because of the periodicity of the waveforms, discontinuities are now located at the block boundaries only. The discontinuities result from two causes: the discontinuities in the amplitude of the contributions of both prototypes, and discontinuities in the phase of the harmonics which make up the periodic functions describing the prototype waveforms. The latter cause of discontinuity disappears when the pitch period is constant, i.e., $p(0) = p(K)$. Thus, if the pitch period is constant, then the discontinuities are solely due to the discrete steps in the scaling factors of Eq. (14).

The blockwise interpolation with time scaling proceeds similar to that with fixed time scale. In the frequency domain, the time scaling corresponds to frequency scaling such that all the harmonics of the periodic signals $u(k,\tau)$, $k = 0, \ldots, K$, are lined up. Thus, this interpolation amounts to linear interpolation of the amplitudes of the harmonics of the speech signal, on a pitch-cycle by pitch-cycle basis.

2.2.3. Continuous interpolation. The previous Subsection showed that blockwise interpolation with time scaling leads to discontinuities at the block boundaries. These discontinuities can be eliminated by replacing the discrete pitch-cycle index k with a continuous function of time, the continuous pitch-cycle index $\kappa(t)$. Thus, the instantaneous pitch period evolves now according to

$$\mathrm{p}(t) = p(\kappa(t)) = \frac{(K - \kappa(t))}{K}\,\mathrm{p}(t_0) + \frac{\kappa(t)}{K}\,\mathrm{p}(t_K), \quad 0 \le \kappa(t) < K, \quad (15)$$

where $\mathrm{p}(t)$ represents the instantaneous pitch period as a function of time t. Time, t, and the instantaneous pitch-period index, $\kappa = \kappa(t)$, are related according to

$$t = t_0 + \int_0^{\kappa} p(\kappa')\,d\kappa' = t_0 + \mathrm{p}(t_0)\kappa + (\mathrm{p}(t_K) - \mathrm{p}(t_0))\,\frac{\kappa^2}{2K}, \quad 0 \le \kappa(t) < K, \quad (16)$$

which is the equivalent of Eq. (2). The inverse relationship is

$$\kappa(t) = \int_{t_0}^{t} \frac{1}{p(t')}\,dt' = \begin{cases} \dfrac{-Kp(t_0) + \sqrt{K^2p^2(t_0) + 2K(p(t_K) - p(t_0))(t - t_0)}}{p(t_K) - p(t_0)}, & p(t_K) \neq p(t_0) \\ \dfrac{t - t_0}{p(t_0)}, & p(t_K) = p(t_0). \end{cases} \quad (17)$$

The length of the interpolation interval is still $t_K - t_0 = (K/2)(p(t_0) + p(t_K))$.

The periodic function of Eq. (12) is now a continuous function of the pitch-period index $\kappa(t)$ and is denoted $u(\kappa(t), \tau)$. Thus, the two periodic functions defining the prototype waveforms at the endpoints of the interpolation interval (t_0 and t_K) are $u(0, p(t_0)\tilde{\tau})$ and $\bar{u}(K, p(t_K)\tilde{\tau})$. The alignment procedure is then

$$\tilde{\xi}_K = \underset{\xi_K}{\operatorname{argmin}}\, D(u(0, p(t_0)\tilde{\tau}), \bar{u}(K, p(t_K)(\tilde{\tau} - \tilde{\xi}_K))). \quad (18)$$

The blockwise interpolation of Eq. (14) is easily modified to continuous interpolation. The *instantaneous excitation pitch-cycle waveform* at κ is

$$u(\kappa, \tau) = u(\kappa, p(\kappa)\tilde{\tau}) = \frac{(K - \kappa)}{K} u(0, p(0)\tilde{\tau}) + \frac{\kappa}{K} u(K, p(K)\tilde{\tau}), \quad 0 \leq \kappa < K. \quad (19)$$

Equation (19) ensures continuity of the magnitude of the waveforms over time. The excitation waveform, $x(t)$, on the interpolation interval is now obtained by concatenation of infinitesimal segments of the instantaneous excitation pitch-cycle waveforms. Consider a concatenation of sections of length $dt = p(\kappa)d\kappa$. Over this infinitesimal interval, the pitch-cycle phase $2\pi\tilde{\tau}$ increased by $2\pi d\tilde{\tau} = (2\pi/p(t))dt$. This shows that, for a signal with continuous phase, we have that $\tilde{\tau} = \kappa(t) + \text{constant}$. Setting the constant equal to zero we have

$$x(t) = u(\kappa(t), p(\kappa(t))\kappa(t)). \quad (20)$$

In Eq. (20) the first argument determines the excitation waveform, and the second argument determines the point of this waveform to be used at $x(t)$. It is convenient to choose the interpolation interval to be an integer number of pitch periods. An example of the resulting interpolation is shown in Fig. 4. The time scale of the figure and the prototype excitation waveforms are identical to those of Figs. 2 and 3. Figure 4 displays an integer number of pitch cycles, starting from the center of the left prototype waveform. For the case shown where $K = 4$, additional prototype waveforms must be known for interpolation outside this range.

Thus, the continuous interpolation algorithm proceeds as follows. The synthesizer has given $p(t_0)$, $p(t_K)$, $u(0, \tau)$, $\bar{u}(K, \tau)$, t_0, and a "desired" endpoint t_{update}. First alignment is obtained with Eqs. (18) and (6) (again, $\xi_K = p(t_K)\tilde{\xi}$). Equation (17) is used to compute $\kappa(t_{update})$. If an integer K is desired, then $\kappa(t_{update})$ is rounded to an integer, K, and t_K is obtained with Eq. (16). To create a sampled output signal, we compute κ for each sample point with Eq. (17). Then we compute the appropriate point of the instantaneous excitation pitch-cycle waveform with (20) and (19). The continuous interpolation described by Eqs. (16)–(20) corresponds to continuous, linear interpolation of the amplitudes of the harmonics of the excitation signal.

2.3. *Adjustment of the Degree of Periodicity*

Usage of the basic PWI discussed in the previous subsection produces excellent quality for speech sections which are highly periodic. However, in other speech signals the reconstructed signal may at times sound somewhat buzzy or reverberant. These distortions are the result of too much periodicity and periodic noise, respectively. In this section we will discuss how additional features of the PWI coder can eliminate these distortions. To this purpose, it is necessary to first define appropriate distortion measures for the

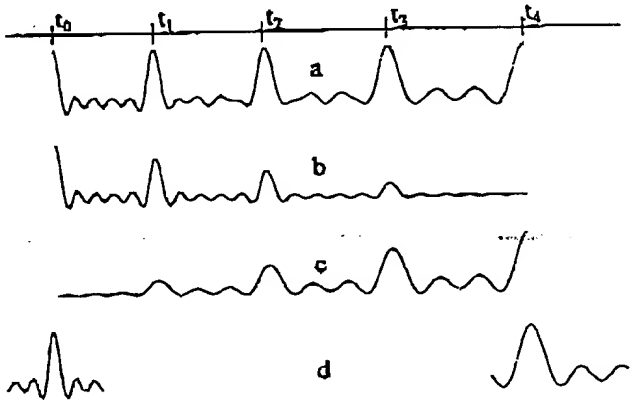


FIG. 4. Continuous interpolation: (a) the interpolated excitation waveform, (b) the contribution from the left prototype excitation waveform, and (c) the contribution of the right prototype excitation waveform. The prototype excitation waveforms are shown in (d).

prototype waveforms. These criteria can also be used for the alignment procedures necessary for interpolation, as described in Subsection 2.2, and the quantization process, which will be discussed in Section 3.

2.3.1. Definition of a distortion measure for prototype waveforms. Consider two excitation waveforms, as defined by Eq. (4), $u(k,\tau)$ and $v(l,\tau)$, and pitch period $p(k)$ and $p(l)$. Further, let $h(t)$ denote the impulse response of a weighting filter. The function of this filter is to spectrally weight the excitations, such that the difference between the filtered prototype excitation waveforms is perceptually relevant. The filter adds the formant structure of the speech signal, but in a deemphasized form. (The deemphasis reflects spectral masking of the human auditory system [13].) The spectral weighting is similar to the weighting used during the search of the codebook in the CELP algorithm [1].

We now define two useful distortion measures between the spectrally weighted excitation waveforms. The first distortion measure is a distortion measure that is invariant with the energy of the individual waveforms, while the second is not. Consider the following, normalized error energy between $u(k,\tau)$ and $\lambda v(l,\tau)$ (λ is an as yet undefined scaling factor),

$$\mathrm{D}_{\lambda}(u(\tau), v(\tau)) = \lim_{P\to\infty} \frac{\int_{-P}^{P} (h(\tau)*u(\tau) - \lambda h(\tau)*v(\tau))^2 d\tau}{\int_{-P}^{P} (h(\tau)*u(\tau))^2 d\tau}, \quad (21)$$

where $*$ denotes convolution. For $\lambda = 1$ this equation provides the second distortion measure. The distortion measure invariant with the energies of the vectors is obtained by determining the value for λ that minimizes the distortion measure of Eq. 21. In that case we obtain

$$\mathrm{D}_{\mathrm{opt}}(u(k,\tau), v(l,\tau)) = 1 - \lim_{P\to\infty} \frac{\left(\int_{-P}^{P} h(\tau)*u(k,\tau)h(\tau)*v(l,\tau)d\tau\right)^2}{\int_{-P}^{P} (h(\tau)*u(k,\tau))^2 d\tau \int_{-P}^{P} (h(\tau)*v(l,\tau))^2 d\tau}. \quad (22)$$

Note that Eq. (22) is symmetric in $u(k,\tau)$ and $v(l,\tau)$; it provides a normalized, energy-invariant, and symmetric measure for the differences between these two excitation waveforms.

The definition of the distortion measures of Eqs. (21) and (22) was aimed mainly at the nonperiodic waveforms of Eq. (4). For the case that the prototype excitation waveforms are assumed to be periodic (as in Eq. (12)), and are time-scaled during interpolation, it is useful to define distortion criteria which operate on the difference between excitation waveforms of normalized pitch. Because the waveforms are periodic, it is convenient to evaluate the distortion measure in the frequency domain. It is then also possible to restrict the criterion to a particular freuency band, or to add frequency-domain weighting. Let $H(\omega)$ be the Fourier transform of $h(\tau)$ and let $U_m(k)$ and $V_m(l)$ be the complex Fourier series coefficients of $u(k,\tau)$ and $v(l,\tau)$, respectively. In general, the pitch periods of $u(k,\tau)$ and $v(l,\tau)$, denoted by $p(k)$ and $p(l)$, are not identical. The energy-invariant distortion measure is

$$\tilde{\mathrm{D}}_{\mathrm{opt}}(u(k,\tau), v(l,\tau), m_0, m_1) = 1 - \frac{\left(\mathrm{Re}\left\{\sum_{m=m_0}^{m_1} U_m^*(k)\left|H\left(\frac{2\pi m}{p}\right)\right|^2 V_m(l)\right\}\right)^2}{\sum_{m=m_0}^{m_1}\left|U_m(k)H\left(\frac{2\pi m}{p}\right)\right|^2 \sum_{m=m_0}^{m_1}\left|V_m(l)H\left(\frac{2\pi m}{p}\right)\right|^2}, \quad (23)$$

where the tilde indicates that we are dealing with pitch-normalized periodic waveforms, and where p is an appropriate time scaling for the filter. $\tilde{\mathrm{D}}_1(u(k,\tau), v(l,\tau), m_0, m_1)$ is defined similarly. Equation (23) compares the prototype waveforms on a harmonic-by-harmonic basis, starting with harmonic m_0 and ending with harmonic m_1. By setting the time scaling of the weighting filter to p, we have effectively time-scaled the pitch periods of both prototype excitation waveforms to this value, and then applied the spectral weighting. Thus, if $u(k,\tau)$ is the unquantized residual, and $v(l,\tau)$ is some codebook entry, which is evaluated for its match, then it is best to choose $p = p(k)$. If we want to compare two excitation waveforms at different times in the same residual signal (as in Subsection 2.3.2), then it is reasonable to choose $p = (p(k) + p(l))/2$ when $p(k)$ and $p(l)$ are close (the comparison will become less useful when $p(k)$ and $p(l)$ are very different).

2.3.2. The signal-to-change ratio. In the context of the PWI coding method, it is convenient to define a measure of the degree of periodicity, the signal-to-change ratio (SCR). The SCR is simply the inverse of the distortion between two prototype waveforms of the same signal, as measured with the distortion measure of Eq. (22) or (23).

For coding purposes, we describe the periodicity of the speech signal with a long-term and a short-term SCR. We define the long-term SCR as the SCR between prototype waveforms separated by 20–30 ms. This time interval is made to coincide with the update rate of the coding system (the rate at which a proto-

type is extracted from the speech signal). The short-term SCR is defined as the SCR between successive pitch cycles.

In PWI, lowering of the long-term SCR for voiced speech segments results in a reverberant quality. If both short-term and long-term SCR are too high, a buzzy quality emerges. A short-term SCR lower than that of the original signal results in a noisy character. Thus, maintaining the SCR of the original signal is essential for good speech quality. When the speech signal is highly periodic, and the prototype waveforms are not quantized, the interpolation methods discussed in Subsection 2.2 will do this, resulting in high-quality speech.

By adjusting the magnitude of the difference in waveform shapes between successively transmitted prototypes, the long-term SCR can be controlled. The SCR on the original prototypes (which would be used in an unquantized system) is measured, and the SCR of the quantized system is constrained to be not larger than this value. Constraining the long-term SCR results in a larger difference between original and reconstructed speech waveforms, but the reconstructed speech increases in perceptual quality, by removing the reverberation.

When constraining the SCR of reconstructed speech, it is important to consider the effect on the individual frequency bands separately. If a single SCR for the entire speech bandwidth is constrained to be identical to that of the original, then the SCR for the higher frequency bands is usually too high. As a result, the suppression of reverberation by maintaining a constant long-term SCR over the entire speech bandwidth will often result in a buzzy quality. It is better to suppress reverberation separately in several frequency bands. Note that this is particularly convenient if the extended prototype is assumed to be periodic and Eq. (23) can be used.

For speakers with low fundamental frequency, i.e., with a pitch period near, or exceeding, 20 ms, the original long-term SCR and the prototype description provide sufficient information for high-quality speech synthesis. However, for speakers with a shorter pitch period, the interpolation procedure may introduce too much short-term periodicity. For these speakers (a majority), the short-term SCR of the reconstructed speech is larger than that of the original signal. The short-term SCR can be corrected by adding noise to the reconstructed signal. This simple addition of white noise to the excitation signal is feasible since the noise component of the original signal is generated by turbulence and has no structure.

In practice, maintaining the short-term SCR such that a perceptually good speech quality results is relatively straightforward. Thus, it was found experimentally that, for voiced speech, the noise contribution can be kept constant without audible distortion for most speakers, and only small distortions for the remaining speakers. For this result, a small amount of noise, increasing with frequency, is injected at frequencies beyond 2 kHz.

3. IMPLEMENTATION OF THE PWI CODING ALGORITHM

In this section, practical implementations of the PWI algorithm are discussed. We limit our discussion to the critical issues of extraction, quantization, and interpolation of the prototype waveforms.

3.1. Extraction of Prototype Excitation Waveforms

Two methods for extracting the prototype waveform from the original signal are reported in this section. Both procedures use an initial pitch estimate. We obtained these estimates using the modified autocorrelation algorithm [14, 15]. The first extraction procedure is a development of the pitch-marker algorithm used previously in single-pulse excitation [9], while the second procedure searches for a pitch cycle using a new, so-called maximum-prediction-gain criterion. Both prototype extraction methods require comparable computational effort and have similar robustness. If estimates concerning the level of the SCR of the original speech are required, then the algorithm based on the maximum-prediction-gain criterion provides the added advantage of high temporal resolution.

3.1.1. Prototype excitation waveform extraction based on pitch markers. The purpose of a pitch-marker algorithm is to detect the beginnings of each pitch cycle in voiced speech. Thus, two adjacent pitch markers provide the boundaries of a local pitch cycle. The prototype excitation waveform is extracted with a time resolution equal to that of the sampling rate of the sampled signal.

The pitch-marker algorithm proposed in [9] is based on the fact that good-quality periodic speech can be obtained by exciting an LP synthesis filter with one delta impulse for each pitch cycle. The locations of these delta impulses are the pitch markers. They are determined using a dynamic programming framework employing a cost function that combines several perceptually important criteria, including the mean-squared error between original and reconstructed speech, smoothness of successive pulse amplitudes and pulse intervals, and deviations of the pulse intervals from an initial average pitch estimate. To find the pitch markers the accumulated cost is

minimized over an interval of around 30 ms in duration.

In our discussion of the pitch-marker-based extraction we will focus on a method which is aimed mainly at operation in conjunction with the definition of the extended prototype through zero-padding (Eq. (4)). When using zero-padded prototype waveforms, it is essential to minimize the discontinuities during the interpolation of the prototype excitation waveforms. Thus, the prototype waveforms must be extracted such that the pitch marker is located sufficiently far from its endpoints.

We now discuss a consistent manner of extracting the prototype excitation waveforms. Figure 5 illustrates the definition of prototype excitation waveforms based on pitch markers obtained with the aforementioned method. The speech signal is partitioned into frames of equal length. (At an 8-kHz sampling rate a typical frame length is 200 samples.) Let us denote the pitch-marker locations in sampling units within the current frame as $n_1, \ldots, n_K$ and the first pitch-marker location in the next frame as n_{K+1}. (Note that the subscripts refer here to the pitch markers and are not identical to the pitch-cycle index of the reconstructed signal.) For each frame, we define a prototype excitation waveform for an interval of length $p(K)$ around pitch marker n_K that is limited by the midpoints between the pitch markers at n_{K-1} and n_K, and pitch markers at n_K and n_{K+1}:

$$p(K) = \tfrac{1}{2}(n_K - n_{K-1}) + \tfrac{1}{2}(n_{K+1} - n_K) = \tfrac{1}{2}(n_{K+1} - n_{K-1}). \quad (24)$$

The sampled, unquantized prototype excitation waveform $v(K, nT)$, where T is the sampling period, is obtained by multiplying the LP residual $e(nT)$ with a rectangular window of length $p(K)$,

$$v(K, nT) = e(nT + \tfrac{1}{2}(p(K) + n_K + n_{K-1})T) \times \Xi(p(K), nT), \quad (25)$$

with $\Xi(p(K), nT)$ as defined in Eq. (11).

The pitch-marker method can also be used to obtain prototype waveforms which can be extended periodically (as in Eq. (12)). If the square window of Eq. (25) is replaced by an asymmetric, tapered window, extending from n_{K-1} to n_{K+1}, with its maximum at n_K, then the Fourier transforms of the resulting extended, windowed prototype waveforms can be used to define a periodic prototype waveform.

3.1.2. Prototype excitation waveform extraction based on maximizing the prediction gain. The present procedure extracts a pitch-cycle waveform, which is continuous when extended periodically (Eq. (12)). For satisfactory results, it is necessary to increase the time resolution beyond the usual 8-kHz sampling rate of the speech signal. Similar benefits of increased resolution were earlier observed for the pitch description of CELP [16]. However, in contrast to current implementations of CELP, the PWI method requires the precise pitch period only for obtaining an accurate description of the prototype and the degree of periodicity; a coarsely quantized version can be used for transmission.

The waveform is extracted with what will be referred to as the maximum-prediction-gain criterion: given a starting point of a speech interval, find the interval length which results in maximum short-term prediction gain of the periodic signal obtained by repeating the interval. In general, if one repeats an interval of arbitrary length, a discontinuity will exist at the boundary points (where the left side of the original interval meets the right side), as is illustrated in Fig. 6. The periodic signal cannot be predicted across these boundaries, and the prediction gain will be lower as a result. However, when the interval length is chosen to be equal to the pitch period, i.e., if the periodicity of the signal equals that of the original speech signal, then the periodic signal can be predicted across these boundaries, and the prediction gain will be high.

One method for implementing the maximum-prediction-gain criterion is to obtain a band-limited Fourier series for each candidate interval. Let p denote the candidate interval length in sampling-period units (p is, in general, not integer). A periodic signal, band-limited at half the sampling frequency, can be represented by a Fourier series with j harmonics, where $j \leq p/2 < j + 1$. Thus, we can fit exactly the

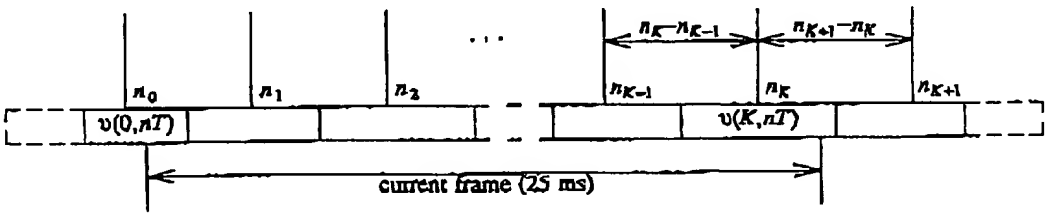


FIG. 5. Extraction of prototype excitation waveforms based on pitch markers.

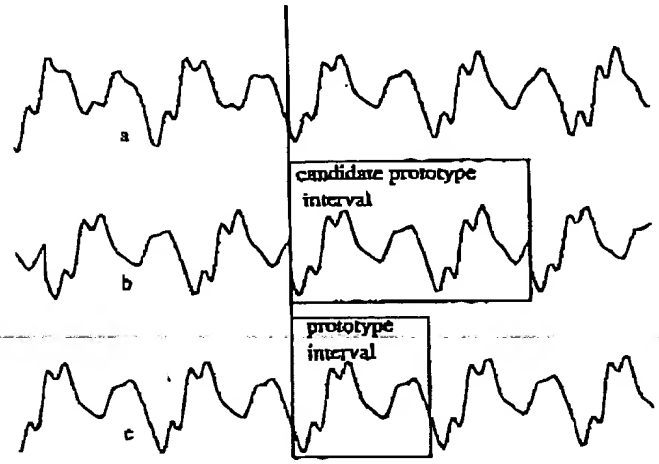


FIG. 6. The maximum-prediction-gain criterion: (a) original speech signal, (b) repetition of an arbitrary speech segment, and (c) repetition of exactly one pitch period.

Fourier series to the speech samples if we center the trial interval of length p at an existing sample. The procedure is best interpreted as a band-limited interpolation for a nonuniformly sampled periodic signal (the sampling pattern being the same for each period, consisting of i identical sampling intervals, and one different sampling interval). Upon obtaining a Fourier series of M harmonics, the autocorrelation function can be computed for all desired lags. Given these autocorrelation values, the short-term predictor coefficients for a periodic signal sampled at the original sampling rate can be computed. From these predictor coefficients the prediction gain can be obtained. Other, more efficient methods of pitch-cycle extraction using the maximum-prediction-gain criterion will be discussed elsewhere.

The maximum-prediction-gain procedure results in a continuous, periodic, band-limited signal, of pitch period $p(K)$, represented by a finite Fourier series. Let us denote the extracted prototype $s(\tau)$ as

$$s(\tau) = \sum_{m=0}^{M} \left[A_m \cos\left(\frac{2\pi m \tau}{p(K)}\right) + B_m \sin\left(\frac{2\pi m \tau}{p(K)}\right) \right], \quad (26)$$

where we included the parameter B_0 for convenience of notation only. The number of harmonics M is a function of the pitch period, $p(K)$, and the cut-off frequency (usually the Nyquist frequency of the sampled speech signal).

The prototype excitation waveform is obtained by filtering a sampled version of the periodic speech signal (assumed to be band-limited) with the digital LP filter with filter coefficients $a_0, a_1, \ldots, a_N$, followed by ideal low-pass filtering at the Nyquist frequency,

$$\tilde{u}(K, \tau) = \sum_{m=0}^{M} \sum_{n=0}^{N} a_n \left(A_m \cos\left(\frac{2\pi m(\tau - nT)}{p(K)}\right) + B_m \sin\left(\frac{2\pi m(\tau - nT)}{p(K)}\right) \right), \quad (27)$$

where T is the sampling interval (used for the digital filtering). Equation (27) implies that, if $s(\tau)$ represents the speech waveform at the update time t_{update}, then the Fourier series coefficients of the unaligned excitation waveform, at the endpoint of the interpolation interval, t_K, are given by

$$\tilde{C}_m = A_m \sum_{n=0}^{N} a_n \cos\left(\frac{2\pi m n T}{p(K)}\right) - B_m \sum_{n=0}^{N} a_n \sin\left(\frac{2\pi m n T}{p(K)}\right)$$

$$\tilde{D}_m = A_m \sum_{n=0}^{N} a_n \sin\left(\frac{2\pi m n T}{p(K)}\right) + B_m \sum_{n=0}^{N} a_n \cos\left(\frac{2\pi m n T}{p(K)}\right), \quad (28)$$

where $\tilde{C}_m$ and $\tilde{D}_m$ are the coefficients for the cosine and sine basis functions, respectively. The tilde indicates that the prototype excitation waveform is not aligned with the previous prototype excitation waveform. The extended prototype excitation waveform can be written as

$$\tilde{u}(K, \tau) = \sum_{m=0}^{M} \left[\tilde{C}_m(K) \cos\left(\frac{2\pi m \tau}{p(K)}\right) + \tilde{D}_m(K) \sin\left(\frac{2\pi m \tau}{p(K)}\right) \right], \quad (29)$$

where the dependence of the Fourier series coefficients on the pitch-cycle index was made explicit by giving them the argument $k = K$.

3.2. *Interpolation of the Prototype Waveform*

3.2.1. Interpolation with sampled time-domain description. The sampled time-domain method is aimed at good performance at a minimal computation cost. The time-domain resolution is limited to 8 kHz during the entire interpolation process. The prototype excitation waveform is described by discrete samples obtained with the extraction method of Subsection 3.1.1.

The first step upon obtaining the sampled prototype excitation waveforms is their alignment. An effective procedure for time alignment of the sampled prototype excitation waveforms takes advantage of the implicit knowledge of the location of the pitch markers. The present prototype excitation waveform

$\tilde{u}_K(nT)$ is shifted according to Eq. (6) such that its pitch-marker position is aligned with the one in the (recentered) previous prototype excitation waveform $u_0(nT)$. Then Eq. (5) reduces to

$$\xi_K = \tilde{n}_0 - \tilde{n}_K, \tag{30}$$

where $\tilde{n}_0$ and $\tilde{n}_K$ are the pitch-marker locations with respect to the center of the window of Eq. (25).

With this specification of the alignment procedure, interpolation of the sampled prototype excitation waveforms is fully described by Eqs. (1)–(11) of Section 2. In natural speech, the number of harmonics occasionally doubles or halves. To prevent unnatural warps of the pitch period in these cases, the smaller pitch period is repeated an integer number of times, such that it most closely matches the length of the larger pitch period.

We now define in more detail the distortion criteria, which are used for adjustment of the SCR, and the quantization process (and as an alternative for the alignment procedure of Eq. (30)). The distortion criteria of Eqs. (21) and (22) are appropriate for the time-domain interpolation of sampled waveforms. To describe this distortion criterion for the discrete signal it is advantageous to approximate the all-pole LP-synthesis filter by its truncated impulse response h_0, $h_1, \ldots, h_R$. (A value of around 25 suffices for R.) To take into account the spectral masking of the human auditory system, this impulse response is modified by the perceptual weighting factor γ to account for spectral masking: $h_0, \gamma h_1, \gamma^2 h_2, \ldots, \gamma^R h_R$. (The perceptual weighting factor usually has a value of around 0.8.) The perceptually weighted response y to a vector u describing a sampled version $u(k, nT)$ of the zero-padded continuous prototype excitation waveform is

$$\mathbf{y} = \mathbf{H}\mathbf{u}, \tag{31}$$

where H is

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & \cdots & \cdots & 0 \\ \gamma h_1 & h_0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & h_0 \\ \gamma^R h_R & \gamma^{R-1} h_{R-1} & \cdots & \cdots & \cdots \\ 0 & \gamma^R h_R & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & \gamma^R h_R \end{bmatrix}. \tag{32}$$

Thus, Eqs. (21) and (22), expressed in the time domain, become for the sampled time-domain description (v describes $v(l, nT)$)

$$D_1(u(k, nT), v(l, nT)) = \frac{(\mathbf{u}-\mathbf{v})^T\mathbf{H}^T\mathbf{H}(\mathbf{u}-\mathbf{v})}{\mathbf{u}^T\mathbf{H}^T\mathbf{H}\mathbf{u}}, \tag{33}$$

$$D_{opt}(u(k, nT), v(l, nT)) = 1 - \frac{(\mathbf{u}^T\mathbf{H}^T\mathbf{H}\mathbf{v})^2}{\mathbf{u}^T\mathbf{H}^T\mathbf{H}\mathbf{u}\ \mathbf{v}^T\mathbf{H}^T\mathbf{H}\mathbf{v}}. \tag{34}$$

If u is considered to be the target excitation [4], and v a candidate excitation vector from the codebook, then the distortion measure of Eq. (34) is of the same form as that used in the quantization of the excitation function in CELP (note that $\mathbf{u}^T\mathbf{H}^T\mathbf{H}\mathbf{u}$ is a constant). However, the interpretation of the target excitation differs. In the CELP algorithm, the continuous update allows one to correct errors in the zero-input response in the current frame of the reconstructed signal (the target excitation waveform is not equal to the residual signal). In the PWI method, the target excitation is identical to the actual excitation, because we consider only one pitch cycle per update interval and correction for previous quantization errors is not applicable.

3.2.2. Prototype waveform interpolation based on the Fourier series description. We now discuss a practical interpolation of the continuous interpolation described in Subsection 2.2.3. The periodic waveforms are described with a Fourier series which can be obtained with the procedures discussed earlier. In the implementation of the continuous interpolation method, we take advantage of the fact that linear interpolation of the shape of the prototype excitation is equivalent to a linear interpolation of their Fourier series coefficients.

If the pitch period has changed over the interpolation interval, the number of harmonics of the prototypes representing the endpoints may not be equal. For the prototype with the lesser number of harmonics, the "missing" harmonics have moved beyond the Nyquist frequency and have been removed by the anti-aliasing filter of the analog-to-digital converter. To facilitate interpolation, harmonics of zero amplitude are added to the prototype with the lesser number of harmonics. To prevent unnatural warps of the fundamental frequency during pitch doubling or pitch halving, the prototype waveform with the smaller pitch period is again repeated an integer number of times, such that it most closely matches the length of the larger pitch period. In the present method, this is equivalent to interspacing zero-amplitude harmonics between the original harmonics for the prototype with less harmonics.

Introducing spectral weighting consists of the inverse of the operation described by Eq. (28). If $\tilde{C}_m(k)$ and $\tilde{D}_m(k)$ are the coefficients of the Fourier series of the excitation function representing the center of pitch cycle k, then the coefficients of the spectrally weighted Fourier series, $\tilde{E}_m(k)$ for cosine and $\tilde{F}_m(k)$ sine basis functions, are

$$\tilde{E}_m(k) = \frac{\tilde{C}_m(k)\sum_{n=0}^{N}\gamma^n a_n\cos\left(\frac{2\pi mnT}{p(k)}\right) + \tilde{D}_m(k)\sum_{n=0}^{N}\gamma^n a_n\sin\left(\frac{2\pi mnT}{p(k)}\right)}{\left(\sum_{n=0}^{N}\gamma^n a_n\cos\left(\frac{2\pi mnT}{p(k)}\right)\right)^2 + \left(\sum_{n=0}^{N}\gamma^n a_n\sin\left(\frac{2\pi mnT}{p(k)}\right)\right)^2}$$

$$\tilde{F}_m(k) = \frac{\tilde{D}_m(k)\sum_{n=0}^{N}\gamma^n a_n\cos\left(\frac{2\pi mnT}{p(k)}\right) - \tilde{C}_m(k)\sum_{n=0}^{N}\gamma^n a_n\sin\left(\frac{2\pi mnT}{p(k)}\right)}{\left(\sum_{n=0}^{N}\gamma^n a_n\cos\left(\frac{2\pi mnT}{p(k)}\right)\right)^2 + \left(\sum_{n=0}^{N}\gamma^n a_n\sin\left(\frac{2\pi mnT}{p(k)}\right)\right)^2} \quad (35)$$

where γ is the perceptual weighting factor, discussed earlier in Subsection 3.2.1.

It is convenient to use a vector notation to describe the distortion criteria. The Fourier series coefficients of $u(k, \tau)$ are represented by the vector

$$\mathbf{U} = [C_0(k) + jD_0(k)\ C_1(k) + jD_1(k) \cdots C_M(k) + jD_M(k)]^H, \quad (36)$$

where the superscript H denotes the Hermitian transpose (conjugate transpose). The linear mapping of Eq. (35) can now be described by the following diagonal matrix W:

$$W_{mm} = \frac{\sum_{n=0}^{N}\gamma^n a_n\cos\left(\frac{2\pi mnT}{p(k)}\right) + j\sum_{n=0}^{N}\gamma^n a_n\sin\left(\frac{2\pi mnT}{p(k)}\right)}{\left(\sum_{n=0}^{N}\gamma^n a_n\cos\left(\frac{2\pi mnT}{p(k)}\right)\right)^2 + \left(\sum_{n=0}^{N}\gamma^n a_n\sin\left(\frac{2\pi mnT}{p(k)}\right)\right)^2}. \quad (37)$$

The vector of Fourier series coefficients of the spectrally weighted signal is then WU. The distortion measures can then be written as

$$\tilde{D}_1(u(k,\tau), v(l,\tau)) = \frac{(\mathbf{U}-\mathbf{V})^H\mathbf{W}^H\mathbf{W}(\mathbf{U}-\mathbf{V})}{\mathbf{U}^H\mathbf{W}^H\mathbf{W}\mathbf{U}}, \quad (38)$$

$$\tilde{D}_{opt}(u(k,\tau), v(l,\tau)) = 1 - \frac{(\mathrm{Re}\{\mathbf{U}^H\mathbf{W}^H\mathbf{W}\mathbf{V}\})^2}{\mathbf{U}^H\mathbf{W}^H\mathbf{W}\mathbf{U}\ \mathbf{V}^H\mathbf{W}^H\mathbf{W}\mathbf{V}}, \quad (39)$$

where V is the vector of Fourier series coefficients of $v(l, \tau)$. If only certain frequency bands are to be considered another diagonal weighting matrix can be added to Eqs. (38) and (39). Using either distortion measure (38) or (39), the alignment procedure of Eq. (18) reduces to

$$\tilde{\xi}_K = \underset{\xi'_K}{\mathrm{argmax}}(\mathrm{Re}\{\mathbf{U}^H\mathbf{W}^H\mathbf{W}\mathbf{V}\})^2 = \underset{\xi'_K}{\mathrm{argmax}} \sum_{m=0}^{M} \{(E_m(0)\tilde{E}_m(K) + F_m(0)\tilde{F}_m(K))\cos(2\pi m\tilde{\xi}'_K) + (F_m(0)\tilde{E}_m(K) - E_m(0)\tilde{F}_m(K))\sin(2\pi m\tilde{\xi}'_K)\}. \quad (40)$$

The Fourier series coefficients for the properly time-aligned prototype excitation waveform are now

$$C_m(K) = \tilde{C}_m(K)\cos(2\pi m\tilde{\xi}_K) - \tilde{D}_m(K)\sin(2\pi m\tilde{\xi}_K),$$
$$D_m(K) = \tilde{C}_m(K)\sin(2\pi m\tilde{\xi}_K) + \tilde{D}_m(K)\cos(2\pi m\tilde{\xi}_K). \quad (41)$$

Once two (quantized) aligned prototype excitation waveforms have been created, the interpolation procedure can proceed according to Eq. (19). An insightful expression for the interpolation process is obtained by first introducing the interpolation function $\alpha(t)$:

$$\alpha(t) = \frac{\kappa(t)}{K}. \quad (42)$$

Then, by writing the integration of Eq. (17) explicitly in Eq. (20), we obtain, over the interpolation interval $t_0 \le t < t_K$,

$$x(t) = \sum_{m=0}^{M}[(1-\alpha(t))C_m(K) + \alpha(t)C_m(0)]\cos\left(\int_{t'=0}^{t}\frac{2\pi m dt'}{(1-\alpha(t'))p(0) + \alpha(t')p(K)}\right) + [(1-\alpha(t))D_m(K) + \alpha(t)D_m(0)]\sin\left(\int_{t'=0}^{t}\frac{2\pi m dt'}{(1-\alpha(t'))p(0) + \alpha(t')p(K)}\right). \quad (43)$$

Note that interpolation of the pitch periods as described in this paper generally leads to a reconstructed speech signal that is not synchronous with the original speech. However, this has no effect on the perceived speech quality. If pitch-synchronous output speech is desired one may apply prototype waveform interpolation without interpolating the pitch periods but using the periods as defined by the pitch markers. In this case, all pitch markers of a frame must be encoded and transmitted which increases the total bit rate by about 760 b/s compared to pitch interpolation [9]. (If the synchronicity is desired only at the endpoint of the interpolation interval, then the pitch period $p(K)$ can be suitably adjusted to achieve this. Additional information is required to prevent propagation of errors in this case.)

Figure 8 compares the narrowband log-magnitude spectra of the original speech of a female speaker with that reconstructed with PWI at a bit rate of 1.7 kb/s, without inclusion of the short-term SCR correction. The spectra were computed by applying a discrete Fourier transform to a 40-ms speech segment. Both signals are band-limited to the frequency range 0.2–3.8 kHz. The envelope and the harmonic structure of both spectra coincide reasonably well. Since the short-term SCR is not considered the reconstructed speech occasionally becomes more periodic in certain frequency intervals than the original speech. This can be seen in Fig. 8 for frequencies above 3 kHz. In addition, the pitch interpolation may sometimes cause relatively large changes from the initial pitch periods as defined by the pitch markers. This may result in a slight frequency shift of the harmonics of the synthetic speech compared to the original speech. Despite the fact that the speech quality improves significantly with increasing bit rate, a comparison of the log-magnitude spectra of reconstructed speech at 1.7 kb/s and higher rates does not reveal significant differences.

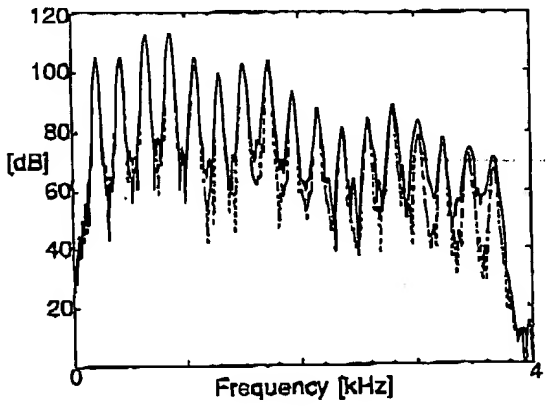


FIG. 8. Comparison of spectra of the original speech (solid line) and the reconstructed speech using PWI at 1.7 kb/s (dashed line).

5. CONCLUSIONS

The interpolation of prototype waveforms leads to excellent voiced speech quality at bit rates between 2.5 and 4 kb/s. Unlike low-bit-rate CELP, the speech is not distorted by background noise. We illustrated the method with several linear-prediction-based techniques. The resulting reconstructed speech signal is, in general, not pitch synchronous with the original signal but displays a similar waveform. The prototype concept facilitates a relatively straightforward speech synthesis system that generates natural sounding voiced speech with a purely harmonic description of the excitation signal, and linear interpolation over 20- to 30-ms intervals. The interpolation procedure facilitates generation of a high level of periodicity, even at low bit rates. Simple procedures can be applied to control the degree of periodicity. Although the PWI method, similarly to single-pulse coding, is currently aimed at voiced speech only, it combines easily with other linear-prediction-based coders such as CELP, which can be used for unvoiced speech.

When the present method is compared to sinusoidal speech coding algorithms, it should be noted that the present method does not require encoding of frequency offsets, birth and death of harmonics (except for pitch doubling or halving), and polynomial interpolation. The phases of all harmonics of the excitation signal are implicitly encoded by a quantized vector describing the prototype excitation waveform. For voiced speech, the low update rate of the pitch parameter and the inclination toward periodicity are the main advantages when compared to current CELP coders.

Although this was not the primary goal of the PWI analysis–synthesis system, it can be used for time and pitch-period scaling. By simply changing the distance over which the prototypes are interpolated, a time-scaled signal is obtained. This was found to work extremely well over a large range of scaling values. Pitch scaling is similarly straightforward. By changing the pitch in the synthesizer by the desired amount, natural sounding speech with a modified pitch is produced.

REFERENCES

1. Atal, B. S., and Schroeder, M. R. Stochastic coding of speech at very low bit rates. *Proc. Int. Conf. Comm., Amsterdam*, 1984, pp. 1610–1613.

2. Atal, B. S., and Hanauer, S. L. Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.* 50, (1971), 637–655.

3. Singhal, S., and Atal, B. S. Improving performance of multi-pulse LPC coders at low bit rates. *Proc. Int. Conf. Acoust. Speech and Signal Process., San Diego,* 1984, pp. 1.3.1–1.3.4.

4. Kleijn, W. B., Krasinski, D. J., and Ketchum, R. H. Improved speech quality and efficient vector quantization in SELP. *Proc. Int. Conf. Acoust. Speech and Signal Process., New York,* 1988, pp. 155–158.

5. Kroon, P., and Atal, B. S. On improving the performance of pitch predictors in speech coding systems. In *Advances in Speech Coding* (B. S. Atal, V. Cuperman, and A. Gersho, Eds.), Kluwer Academic, Dordrecht, Holland, 1991, pp. 321–327.

6. Shoham, Y. Constrained-excitation coding of speech at 4.8 kb/s. In *Advances in Speech Coding* (B. S. Atal, V. Cuperman, and A. Gersho, Eds.), Kluwer Academic, Dordrecht, Holland, 1991, pp. 339–348.

7. Ono, S., and Ozawa, K. 2.4 kbps pitch prediction multi-pulse speech coding. *Proc. Int. Conf. Acoust. Speech and Signal Process., New York,* 1988, pp. 175–178.

8. Atal, B. S., and Caspers, B. E. Beyond multipulse and CELP: Towards high quality speech at 4 kb/s. In *Advances in Speech Coding* (B. S. Atal, V. Cuperman, and A. Gersho, Eds.), Kluwer Academic, Dordrecht, Holland, pp. 191–201.

9. Granzow, W., and Atal, B. S. High-quality digital speech at 4 kb/s. *Proc. Global Telecomm. Conf. (GLOBECOM),* 1990, pp. 941–945.

10. Roucos, S., and Wilgus, A. High quality time-scale modification for speech. *Proc. Int. Conf. Acoust. Speech and Signal Process., Tampa,* 1985, pp. 493–496.

11. Charpentier, F., and Stella, M. G. A diphone synthesis system using and overlap-add technique for speech waveforms concatenation. *Proc. Int. Conf. Acoust. Speech and Signal Process., Tokyo,* 1986, pp. 2015–2018.

12. Quatieri, T. F., and McAulay, R. J. Phase coherence in speech reconstruction for enhancement and coding applications. *Proc. Int. Conf. Acoust. Speech and Signal Process., Glasgow,* 1989, pp. 207–210.

13. Atal, B. S., and Schroeder, M. R. Predictive coding of speech signals and subjective error criteria. *IEEE Trans. Speech Signal Process.* ASSP-27, 3 (1979), 247–254.

14. Sondhi, M. M. New methods of pitch extraction. *IEEE Trans. Audio Electroacoust.* AU-16, 2 (1968), 262–266.

15. Dubnowski, J. J., Schafer, R. W., and Rabiner, L. R., Real-time digital hardware pitch detector. *IEEE Trans. Acoust. Speech Signal Process.* ASSP-24 (1976), 2–8.

16. Kroon, P., and Atal, B. S. On improving the performance of pitch predictors in speech coding systems. *Proc. Int. Conf. Acoust. Speech and Signal Process., Albuquerque,* 1990, pp. 661–664.

17. Paliwal, K. K., and Atal, B. S. Efficient vector quantization of LPC parameters at 24 bits/frame. *Proc. Int. Conf. Acoust. Speech and Signal Process., Toronto,* 1991, pp. 661–664.

Despite the appearance of an integral, Eqs. (42) and (43) can be used directly to implement the algorithm. This is because a slow wandering of the phase of the speech signal waveform is of no practical significance. Thus, numerical integration of (17) is a convenient method of obtaining the argument of the sine and cosine terms. Alternatively, $\kappa(t)$ can be computed with the analytic expression of Eq. (17) and $2\pi\kappa(t)$ can be used as argument for the sine and cosine functions of Eq. (43).

3.3. *Quantization of the Prototype Waveforms*

Since quantization of the coefficients can be performed with standard procedures, we focus here on the quantization of the prototype excitation waveform. In the simplest case, each prototype excitation waveform can be represented as a single impulse, reducing the PWI coding procedure to an LP-based vocoder.

The following method of quantization can be used for all descriptions of the prototype excitation waveform discussed before. The prototype excitation waveforms are encoded differentially. The objective is to approximate the current, aligned prototype excitation waveform, $u(K, \tau)$, with a contribution from the previous, quantized excitation waveform, $v(0, \tau)$, and contributions from one or more codebooks. (The alignment procedure is advantageously performed with respect to the previous, aligned, *quantized* prototype excitation waveform.) The quantized excitation at the endpoint of the interpolation interval, t_K, is

$$v(K, \tau) = \lambda_0 v(0, \tau) + \sum_{l=1}^{L} \lambda_l c^{(l)}_{(q_l)}(\tau), \qquad (44)$$

where $c^{(l)}_{(q_l)}(\tau)$ is the waveform entry with index q_l from codebook l, and the λ_l are scaling factors. The waveform can be quantized by minimizing the distance measure of Eq. (33) for the set of values of $\{\lambda_0, \ldots, \lambda_L, q_1, \ldots, q_L\}$:

$$\{\lambda_0, \ldots, \lambda_L, q_1, \ldots, q_L\} = \underset{\{\lambda'_0, \ldots, \lambda'_L, q'_1, \ldots, q'_L\}}{\operatorname{argmin}} D_1\Big(u(K, \tau), \lambda'_0 v(0, \tau) + \sum_{l=1}^{L} \lambda'_l c^{(l)}_{(q'_l)}(\tau)\Big). \qquad (45)$$

To allow a sequential optimization of these parameters, without reducing the performance, it is useful to orthogonalize the codebook entries of each quantization codebook to the winning entries of the previous optimization stages.

Note that the differential encoding of the prototype excitation waveform is similar in concept to the quantization method of CELP where one first applies a closed-loop pitch prediction [3], or adaptive codebook [4], and then a fixed (stochastic) codebook.

We use this differential coding scheme with two codebooks. The first codebook consists of only single pulses in the time domain, aimed at accurately modeling the pitch pulse (in the sampled-time-domain notation this corresponds to a single nonzero sample; in the Fourier series notation a band-limited pulse is used). Similarly to results for the fixed codebook of CELP [4], we have found that a good choice for the second codebook is one with entries having a sparse set of pulses in the time domain.

We have obtained excellent speech quality using three times 5 bits for the three gain factors, two times 8 bits for the codebooks, and 7 bits for the pitch. At an update rate of 50 Hz, this corresponds to a bit rate of 1.9 kb/s for the excitation signal. It is likely that this bit rate can be reduced when further refinements are introduced. In particular, it is likely that the PWI-coder efficiency can be improved by training the codebooks.

The LP coefficients can be encoded as in other LP-based systems. We have obtained good results with the efficient vector quantization method described in [17].

4. RESULTS

An example of waveforms in PWI speech coding is illustrated in Fig. 7. Figure 7a shows the original speech waveform for a voiced interval and Fig. 7b the pitch markers. The prototype excitation waveforms, which are delimited by the dotted lines, were extracted using pitch markers as described in Subsection 3.1.1. In the example shown, we applied block-wise interpolation with fixed time scale.

The lowest possible bit rate in PWI coding is obtained if each prototype is represented by a single impulse. For this case, PWI reduces to LP vocoding. Figures 7c and 7d show the resulting excitation and reconstructed speech waveforms, respectively. Each prototype waveform is represented by its pitch period $p(K)$, the impulse amplitude, and the set of LP coefficients. By allocating 7 bits for the pitch period, 8 bits for the impulse amplitude, and 24 bits for the vector quantization of the filter parameters, the overall bit rate amounts to 1.7 kb/s for an update interval of 25 ms (3 bits are used for the voiced/unvoiced classification with a resolution of 50 samples). At this bit rate PWI achieves only a vocoder-like speech quality and suffers from some buzziness.

To obtain high-quality speech, the prototype excitation waveforms must be quantized more accurately.

FIG. 7. Waveforms in PWI speech coding with prototype excitation waveform extraction based on pitch markers and with fixed time-scale blockwise interpolation: (a) original speech, (b) pitch markers, (c) interpolated excitation, and (d) reconstructed speech at an overall bit rate of 1.7 kb/s; (e) interpolated excitation and (f) reconstructed speech at an overall bit rate of 2.6 kb/s; (g) interpolated excitation and (h) reconstructed speech obtained with unquantized prototype waveforms.

Figures 7e and 7f show the excitation and reconstructed speech waveforms, at an overall bit rate of 2.6 kb/s. Differential encoding of the prototype excitation waveforms is performed as described in Subsection 3.3. The long-term SCR is maintained identical to that of the original signal by controlling the gains (no subbands were used). The bit allocation is 7 bits for the pitch period, 8 bits for the single pulse location, 8 bits for a fixed codebook component, and 5 bits each for the gain of the previous prototype, the single-pulse amplitude, and the gain of the fixed codebook component. The buzziness present in the reconstructed speech of the 1.7 kb/s example is almost completely removed. It is removed completely by maintaining the short-term SCR similar to that of the original signal. As mentioned before, for the majority of speakers this can be accomplished by injecting a fixed amount of noise for frequencies beyond 2 kHz.

For comparison purposes, Figs. 7g and 7h show the excitation and the reconstructed speech waveforms obtained for the unquantized prototype waveforms. In this case, each prototype excitation waveform is identical to a pitch cycle of the original residual speech signal. The mismatch between the original and reconstructed speech is only due to the interpolation of the intermediate periods between two prototype waveforms.

We have performed formal listening tests in which the voiced sections of the speech reconstructed by several coders was replaced with speech reconstructed by PWI. These tests indicate that the continuous interpolation method, when operating at a rate of 2.6 kb/s, results in a voiced speech quality similar to that of CCITT-standard 32 kb/s ADPCM. When overlap-add procedures are not used, the blockwise interpolation methods perform marginally worse.

US005884253A

United States Patent [19]

Kleijn

[11] Patent Number: 5,884,253

[45] Date of Patent: Mar. 16, 1999

[54] **PROTOTYPE WAVEFORM SPEECH CODING WITH INTERPOLATION OF PITCH, PITCH-PERIOD WAVEFORMS, AND SYNTHESIS FILTER**

[75] Inventor: **Willem Bastiaan Kleijn**, Basking Ridge, N.J.

[73] Assignee: **Lucent Technologies, Inc.**, Murray Hill, N.J.

[21] Appl. No.: 943,329

[22] Filed: Oct. 3, 1997

Related U.S. Application Data

[63] Continuation of Ser. No. 607,295, Jun. 20, 1996, abandoned, which is a continuation of Ser. No. 550,417, Oct. 30, 1995, abandoned, which is a continuation of Ser. No. 179,831, Jan. 5, 1994, abandoned, which is a continuation of Ser. No. 866,761, Apr. 9, 1992, abandoned.

[51] Int. Cl.[6] G10L 5/02

[52] U.S. Cl. 704/223; 704/265

[58] Field of Search 395/2.1, 2.14–2.17, 395/2.28, 2.27, 2.3–2.32, 2.67, 2.71, 2.74, 2.77, 2.78; 704/210, 220

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,624,302	11/1971	Akal	395/2
4,310,721	1/1982	Manley et al.	395/2
4,392,018	7/1983	Fette	395/2.74
4,435,832	3/1984	Asada et al.	395/2.74
4,601,052	7/1986	Sato et al.	395/2.74
4,850,022	7/1989	Honda et al.	704/207
4,910,781	3/1990	Ketchum et al.	704/223
4,989,250	1/1991	Fujimoto et al.	381/38
5,003,604	3/1991	Ozawald et al.	381/38
5,048,088	9/1991	Taguchi	381/38
5,119,424	6/1992	Asakawa et al.	704/208

OTHER PUBLICATIONS

W. Bastiaan Kleijn and Wolfgang Granzow, "Methods for Waveform Interpolation in Speech Coding," Digital Signal Processing, vol. 1, 215–230, Academic Press (1991).

W. B. Kleijn et al. "Improved Speech Quality and Efficient Vector Quantization in SELP", Proc. Int. Conf. ASSP, pp. 155–158 (1988).

S. Ono et al. "2.4 kbps pitch prediction multi-pulse speech coding", Proc. Int. Conf. ASSP, pp. 175–178 (1988).

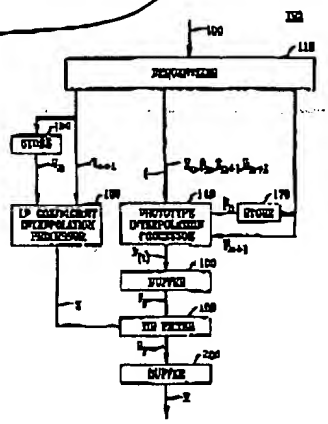
(List continued on next page.)

Primary Examiner—David R. Hudspeth
Assistant Examiner—Donald L. Storm
Attorney, Agent, or Firm—Thomas A. Restaino; Kenneth M. Brown

[57] **ABSTRACT**

A speech coding system providing reconstructed voiced speech with a smoothly evolving pitch-cycle waveform. A speech signal is represented by isolating and coding prototype waveforms. Each prototype waveform is an exemplary pitch-cycle of voiced speech. A coded prototype waveform is transmitted at regular intervals to a receiver which synthesizes (or reconstructs) an estimate of the original speech segment based on the prototypes. The estimate of the original speech signal is provided by a prototype interpolation process which provides a smooth time-evolution of pitch-cycle waveforms in the reconstructed speech. Illustratively, a frame of original speech is coded by first filtering the frame with a linear predictive filter. Next a pitch-cycle of the filtered original is identified and extracted as a prototype waveform. The prototype waveform is then represented as a set of Fourier series (frequency domain) coefficients. The pitch-period and Fourier coefficients of the prototype, as well as the parameters of the linear predictive filter, are used to represent a frame of original speech. These parameters are coded by vector and scalar quantization and communicated over a channel to a receiver which uses information representing two consecutive frames to reconstruct the earlier of the two frames based on a continuous prototype waveform interpolation process. Waveform interpolation may be combined with conventional CELP techniques for coding unvoiced portions of the original speech signal.

10 Claims, 13 Drawing Sheets



Document

Select the documents you wish to save or order by clicking the box next to the document, or click the link above the document to order directly.

save locally as: PDF document include search strategy

previous documents next documents order

document 148 of 190 Order Document

INSPEC - 1969 to date (INZZ)

235

Accession number & update
4065409, B9202-6130-104; 920000.

Title
Methods for waveform interpolation in speech coding.

Author(s)
Kleijn-W-B; Granzow-W.

Author affiliation
AT&T Bell Labs, Murray Hill, NJ, USA.

Source
Digital-Signal-Processing (USA), vol.1, no.4, p.215-30, Oct. 1991.

ISSN
ISSN: 1051-2004.

Publication year
1991.

Language
EN.

Publication type
J Journal Paper.

Treatment codes
T Theoretical or Mathematical; X Experimental.

Abstract
A new method which is positioned between waveform coders and parametric coders is presented. It is based on the assumption that, for voiced speech, a perceptually accurate speech signal can be reconstructed from a description of the waveform of a single, representative pitch cycle per interval of 20-30 ms. The prototype-waveform interpolation (PWI) method retains the natural quality typical of coders which encode the entire waveform, but requires a bit rate close to that of the parametric coders. (17 refs).

Descriptors
encoding; interpolation; speech-analysis-and-processing.

Keywords
speech coding; perceptually accurate speech signal; representative pitch cycle; prototype waveform interpolation; PWI; natural quality; bit rate.

Classification codes
B6130 (Speech analysis and processing techniques).